基于 LOD 的注释服务技术研究*

于倩倩[1,2]　李春旺[1]

[1]（中国科学院国家科学图书馆，北京 100190）[2]（中国科学院研究生院，北京 100049）

[摘要] 本文对基于关联开放数据(LOD)进行的文本、图像和视频等 Web 资源注释服务的相关技术方法进行了梳理和总结，介绍了注释流程中的关联数据查询技术、语义消歧技术、关联扩展技术、关联数据过滤技术和关联模型技术，并提出注释服务应用面临的问题。

[关键词] LOD　注释服务　关联数据

[分类号] G250.7

Study on Technologies of Annotation Service Based on LOD

Yu Qianqian[1,2]　Li Chunwang[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

[Abstract] The annotation service technologies of Web resources such as text, image, video and so on based on LOD are analyzed and summarized, including linked data querying technology, semantic disambiguation technology, relevance expansion technology, linked data filtering technology and relevance model technology, then challenges of annotation service are proposed.

[Keywords] LOD　Annotation service　Linked data

1 背景

注释是附加到其它信息片段的信息[1]。为 Web 资源内容主题如人名、机构名等实体对象或某一领域的主题概念做注释，提供帮助人们理解 web 资源内容的解释、补充片段或元数据信息等即为注释服务。Web 资源种类繁多，包括文本、图像和视频等，对 Web 资源进行注释，一方面可以帮助用户更好地理解知识，另一方面便于用户更准确地搜索到自己所需要的内容。

W3C 启动的关联开放数据(LOD)项目近年来发展极为迅速，截止 2011 年 9 月，LOD 已收录 295 个数据集，提供大约 310 亿个 RDF 三元组以及大约 5.04 亿个 RDF 链接[2]。关联数据的发布与应用为注释服务的发展带来了新的契机，目前越来越多的组织和机构利用关联数据为 Web 资源提供注释服务，其基本原理是针对 Web 资源中的主题，从 LOD 数据集中发现并获取与该主题相关的关联数据信息，帮助用户理解 Web 资源内容以及扩展相关知识。

注释服务是关联参考服务[3]的一种，根据 Web 资源类型的不同，可以将注释服务分为文本注释服务、图像注释服务和视频注释服务等。其中，文本注释服务的相关研究如 Garcia E O 等[4]利用关联数据资源对教学文档中的名词、术语做注释，帮助学生在课程中理解和扩展相关主题的知识；Rusu D 等[5]利用 LOD 中 DBpedia、OpenCyc 和 WordNet 数据集对 web 文本中的主题进行注释，帮助用户理解文本内容等。图像注释服务相关研究如 Sonntag D 等[6]利用 LOD 中 DrugBank、Diseasome 和 DBpedia 三个数据集中的信息为医学图像提供注释功能，支持医生利用扩展信

* 本文系国家社会科学基金资助项目“我国数字图书馆集成融汇方法研究（项目编号：10BTQ004）的研究成果之一”。

息推断可能的疾病并根据病症给出相关的药物信息；Becker C 等[7]利用 DBpedia、GeoNames 和 Freebase 等关联数据集对地图信息做注释，提供用户当前地理位置的背景信息及相关信息，为用户旅行提供导航等。视频注释服务的相关研究如 Haslhofer B 等[8]利用关联数据为视频注释信息做扩展，当用户添加注释信息后，系统自动显示注释信息的相关信息，帮助用户理解视频内容；Ko HG 等[9]利用关联数据对多媒体内容做注释，针对用户观看多媒体内容时输入的关键词，系统自动显示关联数据相关信息，帮助用户消除关键词存在的歧义问题等。

根据资源链接方式的不同，可以将注释服务分为 URI 链接服务、语义扩展服务和元数据添加服务等。其中，URI 链接服务如 Mendes PN 等[10]构建的 DBpedia Spotlight 系统利用 DBpedia URIs 自动注释用户提供的文本片段主题，通过 URI 链接可以发现文本主题的相关信息；Choudhury S 等[11]利用关联数据对 YouTube 视频标签(tag)、用户评论信息做注释，建立视频标签或评论信息到关联数据源中 URI 的链接，通过 URI 链接发现更多的相关信息。语义扩展服务如 Klebeck A 等[12]构建的 Ontos Feeder 利用 DBpedia、Freebase 等关联数据集对网络博客实体对象进行注释，高亮注释实体并通过悬浮窗显示从关联数据集中获取的相关信息；Halb W 等[13]创建的在线内容编辑工具 Link2Wod 关联编辑内容中的术语到关联数据中相关的多媒体信息等，使编辑能更好地控制他们发布的内容等。元数据添加服务的相关研究如 Simon R 等[14]对关联数据为地图提供的注释信息进行保存，在检索时以元数据形式出现，提高检索效果；Virgilio RD 等[15]将识别出的 web 页面实体对象与关联数据相关信息进行链接，以 RDFa 标签形式存储，对 web 页面进行自动注释。

本文对基于 LOD 的 Web 资源注释服务技术方法进行梳理和总结，并对关联数据查询、语义消歧、关联扩展、关联数据过滤、关联模型等注释服务技术进行分析，以便为相关研究提供借鉴。

2 关联数据查询技术

利用关联数据提供注释服务，首先需要将 web 资源主题转换为关联数据的描述形式。关联数据查询即是从 LOD 数据集中获取与 web 资源主题相匹配的关联数据资源的过程。分析已有的注释服务可以发现，关联数据查询技术主要包括 SPARQL 查询、语义网搜索引擎查询和资源匹配等技术方法。

2.1 SPARQL 查询

SPARQL 是一种基于图模式匹配的 RDF 数据查询语言，使用 SPARQL 查询，能够快速获取指定数据源中的相关数据。典型的应用项目如 Latif A 等[16]对计算机科学期刊的作者信息提供的注释服务。

想要获得作者 Arnold Schwarzenegger 的相关信息，首先需要找到与其相匹配的关联数据 URI 如 http://dbpedia.org/resource/Arnold_Schwarzenegger。Latif A 等[16]将 DBpedia 中的 Persondata 数据集通过 RDF Dump 方式下载到本地，获取作者信息，使用 ARC 存储工具构建提供 SPARQL 查询接口的本地三元组存储库。对于用户输入的查询字符串，使用 SPARQL 查询在本地三元组存储库中查找作者的相关信息。SPARQL 查询准确率高，但需要数据源提供 SPARQL 查询端点。此外，SPARQL 查询还可以用来进行关联数据资源遍历和对遍历结果进行过滤。

2.2 语义网搜索引擎查询

语义网搜索引擎如 Sindice[17]、Falcons[18]和 Swoogle[19]等关联数据应用都提供 API 支持搜索关键词。典型的应用项目如 Ko HG 等[9]对多媒体内容如网络电视(IPTV)

进行的注释服务。

Ko HG 等[9]使用 Sindice API 查询用户输入的关键词，选取返回结果中的前 n 个 RDF 结点作为最具代表性的结点。然后通过 SKOS 中的关系属性 skos:broader 和 skos:narrower 比较代表性结点相对的概念层级，选取其中的上位类结点作为匹配结果。语义网搜索引擎可以在整个 LOD 空间中对关联数据进行查询，但是返回的数据质量参差不齐，准确率相对较低。在数据源未知的情况下，可以使用这种方法进行查询。

2.3 资源匹配

资源匹配通过关联数据 URI 解析或关联数据属性信息，获取 Web 资源主题的关联数据描述形式。如将 Web 资源主题词项与 DBpedia 资源的 URIs 匹配或将词项与 DBpedia 资源的 Label 值匹配[20]。

首先，将词项转化为 DBpedia URI 后缀形式（首字母大写或复合词间使用下划线），URI 后缀是去除‘http://dbpedia.org/resource/’之后的字符串；其次将词项与 DBpedia 的 rdf:label 进行匹配，DBpedia 的 labels 是从 Wikipedia 页面的题名创建而来，几乎所有的 DBpedia 资源都提供 rdf:label；然后使用 DBpedia 重定向属性(‘http://dbpedia.org/property/redirect’)获取同义词或字母缩写资源；最后使用 DBpedia 消歧属性(‘http://dbpedia.org/property/disambiguates’)获取多义词资源；如果以上匹配均失败，则将匹配所有以词项作为子串的资源。资源匹配简单、直接，但需要了解数据源使用的词汇表及其表达形式。

3 语义消歧技术

将 web 资源主题与关联数据资源匹配的过程中，由于 web 资源主题本身存在的歧义性，会导致一个 web 资源主题与若干个关联数据资源相匹配。如 Washington 可以指美国第一任总统 George Washington，也可以指城市名 Washington, D.C. 则 Washington 可能与 DBpedia 资源如 dbpedia:George_Washington, dbpedia:Washington,_D.C. 和 dbpedia:Washington_(U.S._state)等相匹配。语义消歧就是从关联数据查询得到的关联数据资源中选择最符合 Web 资源主题上下文的关联数据信息的过程。典型的应用项目如 Mendes PN 等[10]构建的 DBpedia Spotlight 系统使用上下文相似度实现语义消歧;Rusu D 等[5]使用基于 LOD 数据集内容的上下文相似度和基于 LOD 数据集结构的 PageRank 算法实现语义消歧。

3.1 上下文相似度

DBpedia Spotlight[10]是使用 DBpedia URIs 自动注释文本文档的系统。用户提供文本片段(文档、段落、句子),DBpedia Spotlight 注释文本中提及的 DBpedia 资源。首先识别出文本主题上下文即文本片段中该词汇周围的词，如来自同一个段落的词。然后将文本主题与 DBpedia 资源相匹配，找出文本主题在 DBpedia 中的描述形式，对于产生的匹配候选项，将其表示为由来自 Wikipedia 上下文的词项组成的向量，上下文类型有 Wikipedia 页面、消歧页面等，如 Lennon 来自 Wikipedia 文本的词项组成向量{Beatles, McCartney, rock, guitar, ...}。权重计算公式为 TF*ICF，TF 是指词项在上下文中出现的次数，ICF 是逆候选项频率（Inverse Candidate Frequency）。如果 Rs 是文本主题在关联数据中的匹配候选项，$n(w_j)$是 Rs 中与词项 w_j 相关的候选项个数，则

$$\text{ICF}（w_j）= \log \frac{|R_s|}{n(w_j)} = \log|R_s| - \log n(w_j) \quad 公式1$$

最后使用余弦相似度计算文本主题上下文与 DBpedia 中匹配候选项的 Wikipedia 上下文相似度，选择相似度值最大的候选项作为语义消歧项。DBpedia Spotlight 在语义消歧方面取得了较好的应用效果，但目前只能识别 DBpedia 数据集资源，具有一定的局限性。

Rusu D 等[5]使用 LOD 数据集中资源的文本定义，如在 DBpedia 中，符合人们阅读习惯的资源描述定义为 rdfs:comment，对资源的描述类似于摘要。如果文本主题上下文与关联数据候选资源的描述重叠程度越高，则认为两个资源的相近程度越高。将文本主题上下文和候选资源的描述分别定义为 A 和 B 两个词袋模型，使用余弦相似度计算这两个词袋模型的重叠程度，最后选择余弦相似度值最高的候选资源作为语义消歧项。很多 LOD 数据集如 DBpedia、Freebase、OpenCyc 和 WordNet 等都具有资源的文本定义，使用这些数据集进行注释服务时，可以参考这种方法进行语义消歧。

3.2 基于关联数据的 PageRank 算法

PageRank[21]是对网页结构图顶点进行排序的算法，用来标识网页的重要性。LOD 数据集也存在图结构，通过资源之间的关联关系如实例和类之间通过 rdf:type 连接、类和其父类之间通过 rdfs:subclassOf 连接等构成图结构。将 PageRank 算法应用于 LOD 数据集，首先构建 LOD 数据集图 G(V,E)，V 代表数据集的资源，E 代表资源之间的关系，然后识别文本片段中待注释词汇集与关联数据匹配的所有候选资源，即在数据集中，待注释词汇是其 rdfs:label 值的资源。如果图顶点 V 不是候选资源，则将其初始化为 0，否则将其初始化为 1/R，R 是待注释词汇集与关联数据匹配的所有候选资源数量。关联数据资源 i 的 PageRank 值计算公式为，

$$\text{PR}[V_i] = \frac{1-D}{N} + D \bullet \sum_{V_j \in InEdges(V_i)} \frac{PR[V_j]}{OutEdges(V_j)}$$

其中 N 代表图中所有顶点数，调节因子 D=0.85。迭代计算各结点的 PageRank 值，直到图中同一结点两次计算的 PageRank 值相差小于 10^{-15}。最后选择每个待注释词汇的关联数据候选资源中 PageRank 值最高的资源作为其语义消歧项。将 PageRank 算法引入到关联数据中，充分利用了关联数据集的结构信息。对于资源间关联关系较丰富的数据集可以使用这种方式进行语义消歧。

此外，Ludwig N 等[22]结合标签上下文和关联数据中候选资源上下文的共现分析以及实体间关系的链接图分析对用户添加的视频标签进行消歧，Garcia-Silva A 等[23]使用 DBpedia 基于上下文相似度对 web 资源如文本、图像和视频等用户添加的标签进行词义消歧。

4 关联扩展技术

关联扩展是通过 Web 资源主题在关联数据中的描述形式，进一步获取与 Web 资源主题相关的关联数据信息的过程。关联扩展是注释服务过程中的重要环节，是关联数据资源发现获取的重要途径。从已有的注释服务可以看出，选择 LOD 数据集中的相关属性获取关联数据资源是关联扩展的一种方式，还可以在遍历属性的基础上通过结点间的相似性计算选取相关的关联数据资源进行注释。

4.1 关联数据属性选择

关联数据依据 RDF 模型的“资源-属性-属性值”的形式进行表达，通过关联数据中的属性，可以直接获取相关的关联数据资源。但关联数据中的属性繁多，如在 DBpedia 本体中，描述人物的相关属性就有 350 多个[24]，因此，需要选择合适的关联数据属性，以提高注释服务的精准性。典型的研究项目如 Waitelonis J 等[202526]选择关联数据中的重要属性并对其进行排序，依据这些属性搜索相关资源对 Yovisto 学术视频元数据进行注释；Stan J 等[16]使用关联数据中的三种关联关系进行语义扩展，对社交论坛用户发布信息中的关键词和实体对象进行注释。

Waitelonis J 等[202526]认为相关资源的重要程度可以根据属性的重要程度进行衡量，提出启发法对 DBpedia 中的重要属性进行排序。相关属性排序如下：（1）RDF 属性频次。具有 rdf:type 或 skos:subject 的实体的属性频次越高，属性越重要。如果一个实体属于几个分类，则相同属性发生次数加和。（2）具有相同 rdf:type 资源的属性。如果两个资源的 rdf:type 资源是相同的，则连接这两个资源的属性是重要的。（3）预定义类型的属性。如 dbpedia:Event 和 dbpedia:Place，且 dbpedia:Event 重要度高于 dbpedia:Place。（4）双重链接属性。资源间链接属性不同，却是互相指向。（5）dbpedia:disambiguates 属性。（6）dbpedia:wikilink 属性。具有双向 wikilinks 链接的资源比只有单向 wikilinks 链接的资源重要度更高。（7）Wikilinks 入链。（8）List 属性。指那些 URI 后缀以 List_of_结尾的资源，如 dbpedia:List_of_Nobel_laureates。（9）skos:subject 属性。（10）rdf:type 属性。（11）label 子串。实体映射到关联数据时作为子串匹配的资源。

Stan J 等[27]使用的关联关系如下：首先是层次链接（hierarchical links），关联到上位类概念，以属性 subject 表示；其次是有相同上位类的邻结点，以属性 isbroaderof(subject(c))表示；最后是与起始概念直接关联的概念，如 Clint Eastwood 是 Gran Torino 的导演。

综上，基于属性进行资源遍历，需要对关联数据集使用的本体及词汇表等比较了解，选取的属性不同，获取的关联数据资源也不尽相同。使用关联数据属性进行资源遍历简单快捷，获取的资源准确率高，但需要对重要属性进行筛选。

4.2 结点相似性计算

在遍历关联数据属性的基础上，通过计算结点间的相似性选择相关的关联数据资源。典型的应用项目如 Mirizzi R 等[28293031]开发的 SWOC、Not Only Tag 和 LEO 系统，使用关联数据和外部数据源对 IT 领域的词汇概念做注释，这三个系统的后台都是 DBpediaRanker 系统，其主要功能是计算 DBpedia 结点间的相似度；Stankovic M 等[32]使用 hyProximity 计算 DBpedia 中概念间的邻近关系。

两个研究项目都是首先选择种子概念作为概念扩展的起点，如 DBpediaRanker 由领域专家挑选的数据库和编程语言领域的代表性结点为 PHP、Java、MySQL、Oracle、Lisp、C#和 SQLite。然后使用 skos:subject 和 skos:broader 属性遍历数据集资源，skos:subject 表示某个概念属于某个类，skos:broader 表示某个类属于某个上位类。根据研究领域的不同，遍历深度也不同，如 DBpediaRanker 中遍历深度设为 2。

两个项目的不同之处在于对结点的相似度计算方式不同，DBpediaRanker 系统使用任意两个结点在不同数据源中的相似度权重和作为最终的相似度计算结果，hyProximity 基于遍历结点与种子结点的距离进行计算。

在 DBpediaRanker 系统[28293031]中，对于搜索到的任意两个 DBpedia 资源结点，

使用网络搜索引擎（谷歌、雅虎和必应）、社交标签系统（Delicious）和 DBpedia 数据源，计算其相似度。选择不同的搜索引擎，可以不局限于一种搜索引擎的算法，使用社交标签系统，除了考虑词汇在网页中的流行度，还考虑了词汇在用户间的流行度。在搜索引擎和社交标签系统中，对于搜索到的两个 DBpedia 资源 uri_1 和 uri_2，相似度计算公式如下：

$$sim(uri_1, uri_2, is) = \left(\frac{Puri_1, uri_2}{Puri_1} + \frac{Puri_1, uri_2}{Puri_2}\right)_{is}$$

其中，is 代表数据源，p_{uri1} 和 p_{uri2} 分别代表数据源中包含 uri_1 的 rdfs:label 集词汇和 uri_2 的 rdfs:label 集词汇的网页数，$P_{uri1,\ uri2}$ 代表数据源中同时包含两者 rdfs:label 集词汇的网页数。

在 DBpedia 中，对于资源 uri_1 和 uri_2，一方面考虑 Wikipedia 到 DBpedia 的超文本链接 wikilink 属性。如果资源 uri_1 到 uri_2 有 wikilink 属性，资源 uri_2 到 uri_1 也有 wikilink 属性，则 wikiS(uri_1, uri_2)值为 2；如果只有资源 uri_1 到 uri_2 的 wikilink 属性，或只有 uri_2 到 uri_1 的 wikilink 属性，则 wikiS(uri_1, uri_2)值为 1；如果资源 uri_1 和 uri_2 之间没有 wikilink 属性，则 wikiS(uri_1, uri_2)值为 0。另一方面，检查 uri_1 的 rdfs:label 是否包含在 uri_2 的 dbpprop:abstract 中，反之亦然，假设 n 是资源 label 的个数，m 是 abstract 中包含资源 label 的个数，则 abstractS(uri_1,uri_2) =m/n，其值在[0,1]间浮动。

资源 uri_1 和 uri_2 之间的相似度是上述计算的权重和，公式为：

$$similarity(uri_1, uri_2) = w_{google} * sim(uri_1, uri_2, google) + w_{yahoo} * sim(uri_1, uri_2, yahoo) + w_{bing} * sim(uri_1, uri_2, bing) + w_{delicious} * sim(uri_1, uri_2, delicious) + w_{wikiS} * wikiS(uri_1, uri_2) + w_{abstract} * abstractS(uri_1, uri_2)$$

其中权重 w 均设为 1。与单独使用外部资源、单独使用关联数据文本和链接资源、同时使用这两种资源但外部资源中相似度计算使用共现分析等算法相比，本系统资源间相似度计算方法更能反映两个资源之间的关系，具有明显的优势。在关联数据集资源描述较丰富的情况下，注释服务可以借鉴这种方法进行结点相似度计算。

Stankovic M 等 [32] 基于 DBpedia 结构图的两个主要特征：（1）与起始概念距离越短的概念越相关（2）与若干起始概念邻近的概念比与一个起始概念邻近的概念要更相关，计算结点间的相似性。概念 c 到起始概念集 IC 的 hyProximity 计算公式为：

$$hyP(c, IC) = \sum_{c_i \in IC} \frac{p(c, c_i)}{d(c, c_i)}; \quad p(c, c_i) = e^{-\lambda d(c, c_i)}$$

其中，$d(c,c_i)$是 c 与 c_i 的距离，是 c 与 c_i 共享祖先的最短路径，$p(c,c_i)$对不同的距离赋予不同的权重，以指数形式在距离上减少概念重要性，λ=0.3。使用算法完成 hyProximity 的计算并进行排序，选择计算结果较高的值作为相近概念，算法限于第 3 层。该方法利用了关联数据集的结构信息，在关联数据集结构较丰富的情况下，注释服务可以借鉴这种方法进行结点相似度计算。

5 关联数据过滤技术

从 LOD 数据集中获取的数据资源通常是比较多的，为了保证获取资源的质量，

需要对不相关的资源进行过滤。数据过滤可以在资源发现过程中执行，典型研究项目如 Lama M 等[43334]、Ko HG 等[9]、Stan J 等[27]和 DBpediaRanker 系统[28]对新遍历结点的处理过程；也可以在资源发现获取后执行，如 Stankovic M 等[32]、DBpediaSpotlight 系统[10]和 DBpedia Mobile 应用[7]等对已获取资源的处理方法。

5.1 资源发现过程中的数据过滤

Lama M 等[43334]利用关联数据为教育资源添加注释。首先识别出教育资源的主题词，然后搜索 DBpedia 找到相关主题的资源，使用深度优先算法对 DBpedia 资源进行遍历。对遍历的 DBpedia 属性赋予一定的权重值，如属性 skos:broader 的权重值 wr_{SB} 设为 0.6、属性 rdf:type 的权重值 wr_{RT} 设为 0.9 等，可以人工赋予，也可以自动获得。如果遍历资源为叶子结点，即三元组宾语为文字，则通过如下公式计算叶子结点与主题词的关系，

$$\mu(x) = a \times \sum_{i=1}^{K} (wt_i \times ft_i) + b \times \frac{S_x}{K}$$

其中，K 是相关的教育资源主题词数目，i 是第 i 个相关的主题词， wt_i 是第 i 个主题词的权重，ft_i 是第 i 个主题词在结点 x 中的频次，S_x 是教育资源与结点 x 共有的主题词数。如果遍历资源为分支结点，即三元组宾语为 URI，则通过如下公式计算分支结点与主题词的关系，

$$\mu(x) = \sum_{i=1}^{N_x} wr_i \times \mu(t_i)$$

其中，N_x 是结点 x 的属性数目，i 是第 i 个属性，wr_i 是属性的权重，t_i 是第 i 个属性关联的结点。

Ko HG 等[9]使用语义网搜索引擎的查询响应次数计算结点间的相似度，通过相似度的计算过滤掉不相关的结点。Stan J 等[27]计算扩展集中的每个概念与关联数据中的起始概念摘要的相似度，摘要中含有很多与起始概念相关的关键词，这些关键词可作为起始概念的上下文过滤掉扩展集中不相关的概念。在 DBpediaRanker 系统[28]中，将新遍历的结点与该领域流行度最高的 DBpedia 分类进行相似度计算，计算方法为关联扩展技术中介绍的不同数据源中结点的相似度计算权重和，如果相似度值高于给定阈值，则认为新遍历结点属于给定领域上下文，如果相似度值低于给定阈值，则将其过滤掉。由此可以看出，资源发现过程中的数据过滤方法主要是结点间的相似度计算，与关联扩展中的结点相似度计算方法通用，但应用场景不同，主要取决于不同应用对相似度计算阈值的选择。

5.2 资源获取后的数据过滤

Stankovic M 等[32]对主题概念进行注释，基于 DBpedia 中的概念类型，过滤掉人、公司和书籍等不相关的概念。DBpediaSpotlight 系统[10]基于 DBpedia、Freebase 和 Schema.org 的本体分类层级概念类型或通过 SPARQL 语句对获取的关联数据资源进行过滤。DBpedia Mobile 应用[7]构建基于资源类型、评价等的简单过滤或构建 SPARQL 语句过滤掉不相关的资源。资源获取后的数据过滤主要使用关联数据集资源类型和 SPARQL 语句进行过滤，对于需要注释特定类型主题的资源，可以使用这种方法对获取到的关联数据进行过滤。

6 结语

通过分析基于关联数据进行注释服务的已有研究可以发现，其对 Web 资源的注释

主要是对实体对象或主题概念的注释，如 Yovisto 项目[35]注释的视频元数据包括与视频相关的关键词（从题名、演讲者和描述信息等抽取）以及与视频时间相关的关键词（如使用 OCR 方法从视频中抽取）和用户对视频所做的标签，极少涉及对某个段落主题或整篇文章主题的注释服务。通过关联数据查询技术、语义消歧技术、关联扩展技术和关联数据过滤技术，可以获取与注释对象相关的关联数据资源。此外，关联模型可以对数字文献中实体对象的相关属性及关联关系进行描述，以指导关联信息发现、融合与可视化呈现等操作[3]。如 Latif A 等[36]设计了 CAF-SIAL 概念集成框架，支持不同人物类型如科学家、艺术家等相关信息的集成融汇；刘媛媛等[37]基于已有的本体（如 DBpedia 本体）描述方案设计了数字文献中科研人员、科研机构、研究项目等典型实体对象的关联模型。首先根据关联数据源中已有的本体描述结合用户潜在的信息需求对实体对象的相关属性以及与其它实体的关联关系属性进行遴选，构建关联模型；然后依据预先定义的实体间关联关系对注释服务的检索结果进行查询扩展，从而能获取更多的关联数据资源。使用关联模型技术，可以扩展与优化关联数据的检索结果，但是目前关联模型主要是针对数字文献中的实体对象，具有一定的局限性。笔者将继续对注释服务相关环节的技术方法进行研究，同时关注段落主题或整篇文章主题的注释服务，以期在已有方法的基础上尝试构建基于 LOD 的文献主题注释服务应用，在实践中发现问题和解决问题。错误!未找到引用源。错误!未找到引用源。**参考文献：**

[1] Linked Data and multimedia: the state of affairs
[2] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[3] 刘媛媛，李春旺，黄永文. 基于 LOD 的关联参考服务构建研究[J]. 现代图书情报技术，2011，(06):66-71.
[4] Garcia E O, Vidal J, Lama M, et al. Semantic Annotaion of Education Resources through Linked Data[C]. In: Proceedings of ICWL. 2010:311-320.
[5] Rusu D, Fortuna B, Mladenic D. Automatically Annotating Text with Linked Open Data[C]. In: Proceedings of LDOW 2011, Hyderabad, India.2011.
[6] Sonntag D,Wennerberg P.Applications of an Ontology Engineerging Methodology Accessing Linked Data for Medical Image Retrieval[C].In :Proceedings of Linked AI,Califonia,USA.2010:120-125.
[7] Becker C, Bizer C. Exploring the Geospatial Semantic Web with DBpedia Mobile[J]. Web Semantics: Science, Services and Agents on the World Wide Web. 2009, 7(4):278-286.
[8] Haslhofer B,Momeni E.Augmenting Europeana Content with Linked Data Resources[C].In :Proceedings of I-semantics 2010,Craz,Austria.
[9] Ko H G, Ko I Y. Generation of Semantic Clouds Based on Linked Data for Efficient Multimedia Semantic Annotation[C]. In: Proceedings of ICWE 2011, Paphos, Cyprus. 2011:127-134.
[10] Mendes P N, Jakob M, Garcia-Silva A, et al. DBpedia Spotlight: Shedding Light on the Web of Documents[C]. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), Graz, Austria. 2011.
[11] Choudhury S,Breslin J G,Passant A.Erichment and Ranking of the YouTube Tag Space and Integration with the Linked Data Cloud[C].In:Proceedings of ISWC,Berlin Heidelberg. 2009:747-762.
[12] Klebeck A, Hellmann S, Ehrlich C, et al. OntosFeeder – A Versatile Semantic Context Provider for Web Content Authoring[C]. In: Proceedings of ESWC 2011
[13] Halb W,Stocker A,Mayer H et al.Towards Commercial Adoption of Linked Open Data for Online Content Providers[C].In:Proceedings of I-Semantics,Graz,Austria,2010.
[14] Rainer Simon, Bernhard Haslhofer, Joachim Jung. Annotations, Tags & Linked Data - Metadata Enrichment in Online Map Collections through Volunteer- Contributed Information.
[15] RDFa Based Annotation of Web Pages through Keyphrases Extraction OTM 2011
[16] Latif A, Afzal M T, Ussaeed A, et al. Turning keywords into URIs: simplified user interfaces for exploring linked data[C]. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea. 2009:76-81.
[17] Sindice-The semantic web index[EB/OL]. [2012-05-08]. http://sindice.com/.
[18] Cheng G, Ge W, Qu Y Z. Falcons: Searching and Browsing Entities on the Semantic Web[C]. In Proceedings of the WWW, Beijing, China. 2008.

[19] Cheng G, Ge W, Qu Y Z. Falcons: Searching and Browsing Entities on the Semantic Web[C]. In Proceedings of the WWW, Beijing, China. 2008.
[20] Waitelonis J, Sack H. Towards exploratory video search using linked data[C]. In: Proceedings of the 11th IEEE International Symposium on Multimedia, Washington, DC, USA. 2009:540-545.
[21] Brin S, Page M. Anatomy of a large-scale hypertextual Web search engine[C]. In: Proceedings of the 7th Conference on World Wide Web (WWW), Brisbane, Australia. 1998.
[22] Ludwig N, Sack H. Named Entity Recognition for User-Generated Tags[C]. In: Proceedings of the 22nd International Workshop on Database and Expert Systems Applications (DEXA). 2011:177-181.
[23] Garcia-Silva A, Szomszor M, Alani H, et al. Preliminary Results in Tag Disambiguation using DBpedia[C]. In: Proceedings of the 1st International Workshop on Collective Knowledge Capturing and Representation at K-CAP 2009, Redondo Beach, California, USA. 2009.
[24] DBpedia ontology class Scientist [EB/OL].
[2011-10-08].http://mappings.dbpedia.org/server/ontology/classes/Person.
[25] Waitelonis J, Sack H, Kramer Z, et al. Semantically enabled exploratory video search[C]. In: Proceedings of the 19th World Wide Web Conference (WWW2010), Raleigh, NC, USA. 2010.
[26] Waitelonis J, Sack H. Towards exploratory video search using linked data[J]. Multimedia Tools and Applications, 2012,59(2):645-672.
[27] Stan J, Do V H, Maret P. Semantic User Interaction Profiles for Better People Recommendation[C]. In: Proceedings of ASONAM 2011, Taiwan, China. 2011.
[28] Mirizzi R, Ragone A, Di Noia T, et al. Ranking the linked data: the case of dbpedia[C]. In: Proceedings of ICWE 2010. 2010:337 - 354.
[29] Mirizzi R, Ragone A, Di Noia T, et al. Semantic wonder cloud: exploratory search in dbpedia[C]. In: Proceedings of the 2th International Workshop on Semantic Web Information Management (SWIM 2010). 2010:138-149.
[30] Mirizzi R, Ragone A, Di Noia T, et al. Semantic tags generation and retrieval for online advertising[C]. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM2010), Toronto, Ontario, Canada. 2010.
[31] Mirizzi R, Ragone A, Di Noia T, et al. What if exploratory search and web search meet?[C]. In: Proceedings of the Fourth Ph.D. Workshop in CIKM, PIKM 2010, Toronto, Canada. 2010.
[32] Stankovic M, Breitfuss W, Laublet P. Linked-Data Based Suggestion of Relevant Topics[J]. In: Proceedings of the 7th International Conference on Semantic Systems9(I-SEMANTICS 2011), Graz, Austria. 2011:49-55.
[33] Lama M, Vidal J C, Otero-Garcia E, et al. Semantic Linking of a Learning Object Repository to DBpedia[C]. In: Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies (ICALT 2011). 2011 :460-464.
[34] Vidal J C, Lama M, Otero-Garcia Eet al. An evolutionary approach for learning the weight of relations in linked data[C]. In: Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011). 2011:1002-1007.
[35] Waitelonis J, Nadine L, Sack H. Use What You Have: Yovisto Video Search Engine Takes a Semantic Turn[C]. In: Proceedings of the 5th International Conference on Semantic and Digital Media Technologies(SAMT2010), Saarbrücken, Germany. 2010.
[36] Latif A,Afzal M T,Saeed A U,Hoefler P,Tochtermann K.CAF-SIAL:Concept Aggregation Framework for Structuring Information Aspects of Linked Open Data[C].In:Proceedings of International Conference on Networked Digital Technologies,Ostrava,Czech Repblic.2009:100-105.
[37]刘媛媛.基于 LOD 的关联参考服务构建研究[D].北京：中国科学院国家科学图书馆，2011.

(作者 E-mail: yuqianqian@mail.las.ac.cn)